

### **REMARKS**

This Supplemental Amendment is in response to the Office Action of May 26, 2004, and is further and in addition to the previously-filed Amendment of August 24, 2004. In the interest of clarity, all amendments to the claims in the previously-filed Amendment are included in this Supplemental Amendment, such that the Examiner has a complete set of modified claims on which to act. Therefore, the present Amendment amends claims 1, 14, 18, 20-28, 37 and 39 in accordance with the originally-filed specification and cancels claims 13, 17 and 19. Accordingly, claims 1-12, 14-16, 18 and 20-39 remain in this application, and claims 1, 37 and 39 are in independent form.

Initially, the Examiner is thanked for indicating that the subject matter of claims 22-25 and 31-36 define over the prior art of record. In particular, the Examiner has indicated that these claims are objected to as being dependent upon a rejected base claim, but would be allowable if rewritten in independent form including all the limitations of the base claim and any intervening claims. Further, with respect to claims 22 and 24, these claims determine a distinctiveness score as a function of the frequency of a particular text fragment within a document (or text spans) in the large collection of documents (or text spans). The Examiner also indicated that claim 23 would be allowable if rewritten to overcome the rejections under 35 U.S.C. § 112, second paragraph, set forth in the Office Action.

Claims 15, 19 and 23 stand rejected under 35 U.S.C. § 112, second paragraph, as being indefinite for failing to particularly point out and distinctly claim the subject matter which Applicant regards as the invention. Further, the Examiner has objected to claim 20 under 37 C.F.R. § 1.75 as being a substantial duplicate of claim 19. All of the Examiner's rejections and objections to claims 15, 19, 20 and 23 have been addressed and overcome by the Amendment of August 24, 2004. For a detailed explanation of how these rejections and objections have been overcome, please see the Remarks section of the previously-filed Amendment, which is incorporated herein by reference. Therefore, Applicant respectfully requests withdrawal of these rejections and objections to claims 15, 19, 20 and 23.

Claims 1, 2, 10-21, 26-29 and 37-39 stand rejected under 35 U.S.C. § 103(a) as being obvious over U.S. Patent No. 6,615,209 to Gomes et al. (hereinafter "the Gomes patent"). Further, claims 3-7 and 30 stand rejected under 35 U.S.C. § 103(a) as being obvious over the Gomes patent in view of U.S. Patent No. 6,240,409 to Aiken. Finally, claims 8 and 9 stand rejected under 35 U.S.C. § 103(a) as being obvious over the Gomes patent in view of U.S. Patent No. 6,356,633 to Armstrong. In view of the previously-filed Amendment, the previously-filed Declaration of Mark Kantrowitz and the following remarks, Applicant respectfully requests reconsideration of these rejections.

First, it is noted that the Gomes patent has been used as the primary reference in all the rejections by the Examiner. However, the invention underlying the present application was conceived prior to the filing date of the provisional patent application underlying the Gomes patent, and the present invention was diligently reduced to practice through the filing date of the present application. Accordingly, the Gomes patent does not represent prior art and should be withdrawn from consideration. The underlying facts surrounding this assertion are discussed in great detail in the previously-filed Amendment, as well as the previously-filed Declaration and accompanying documentation. Therefore, these remarks, documents and arguments are incorporated herein by reference and will not be duplicated herein.

Notwithstanding the requisite withdrawal of the Gomes patent as prior art to the present invention, Applicant has further modified various claims in the present application through the foregoing Amendment, including all of independent claims 1, 37 and 39. In particular, independent claim 1 is directed to a computer-assisted method for identifying duplicate and near-duplicate documents in a large collection of documents. This method includes the steps of: initially, selecting distinctive features contained in the collection of documents; then, for each document, identifying the distinctive features contained in the document; and then, for each pair of documents having at least one distinctive feature in common, comparing the distinctive features of the documents to determine whether the documents are duplicate or near-duplicate documents. The distinctive features are text

fragments, which are sequences of at least two words that appear in a limited number of documents in the document collection, and the text fragments are determined to be distinctive features based upon a function of the frequency of a text fragment within a document in the large collection of documents.

Independent claim 39 of the present application, as amended, is an apparatus claim that corresponds to independent claim 1 as discussed above. Specifically, this claim also includes the limitations that: (1) the distinctive features are text fragments, which are sequences of at least two words that appear in a limited number of documents in the document collection; and (2) the text fragments are determined to be distinctive features based upon a function of the frequency of a text fragment within a document in a large collection of documents.

Independent claim 37 of the present application, as amended, is directed to a computer-assisted method for identifying duplicate and near-duplicate text spans in a large collection of text spans. Again, this independent claim has been modified to conform with the limitations set forth in independent claims 1 and 39, and varies from these claims in that the term "text spans" is used in place of "documents".

The Gomes patent is directed to the detection of query-specific duplicate documents. In particular, query-relevant information is used to limit documents for comparison or similarity. Once the document list is condensed based upon this query, specific and query-relevant information or text is extracted from the documents. It is this extracted information or text "snippets" that is used to compare the documents for similarities.

The Aiken patent is directed to a method and apparatus for detecting and summarizing document similarity within large document sets. It appears that the Examiner is using the Aiken patent for its teaching of using such a method to detect plagiarism, copyright infringement, authorship of a document, as well as version and text clustering algorithms and calculating overlapped ratios.

The Armstrong patent is directed to electronic mail message processing and routing for call center response. It appears that the Examiner is using the Armstrong patent for

its disclosure of matching an e-mail message with responses to the e-mail message, and similar electronic mail messaging functionality.

Again, Applicant has already effectively “sworn behind” the Gomes patent. However, even if the Gomes patent is considered prior art, independent claims 1, 37 and 39, as amended, and the methods and apparatus disclosed therein, are distinguishable from the Gomes patent and the remaining prior art of record, whether used alone or in combination. Specifically, the Gomes patent chooses various distinctive features or “snippets” based upon their presence in a query. However, the present invention identifies distinctive features in an entirely different and novel manner. First, the distinctive features are text fragments, which are sequences of at least two words that appear in a limited number of documents (or text spans) in the document (or text span) collection. Importantly, the text fragments are determined to be distinctive features based upon a function of the frequency of a text fragment within a document (or text span) in the large collection of documents (or text spans). Therefore, “distinctiveness” is a function of the frequency of each candidate feature (phrase or text fragment) within each document in the set of documents. The determination of “distinctiveness” or distinctive features based upon a function, as opposed to presence in the query, illustrates the novel concept involved in the present invention. Still further, the Examiner has already indicated that the subject matter of claims 22 and 24 of the present application define over the prior art of record. It should be noted that both of these claims discuss specific functions for determining distinctiveness, so the modifications to independent claims 1, 37 and 39 incorporate, in a general sense, the allowable subject matter of the present application.

For the foregoing reasons, and for the reasons set forth in the previously-filed Amendment, independent claims 1, 37 and 39 are not anticipated by or rendered obvious over the Gomes patent, the Aiken patent, the Armstrong patent or any of the prior art of record, whether used alone or in combination. There is no hint or suggestion in any of the references cited by the Examiner to combine these references in a manner which would render the invention, as claimed, obvious. Reconsideration of the rejection of independent claims 1, 37 and

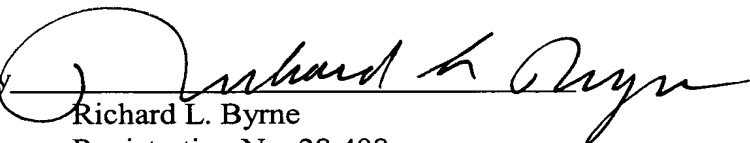
39 is respectfully requested.

Claims 2-12, 14-16, 18, and 20-36 depend either directly or indirectly from and add further limitations to independent claim 1 and are believed to be allowable for the reasons discussed hereinabove in connection with independent claim 1. Further, claim 38 depends directly from independent claim 37 and is believed to be allowable for the reasons discussed hereinabove in connection with independent claim 37. Therefore, for all of these reasons, reconsideration of the rejection of claims 2-12, 14-16, 18, 20-36 and 38 is respectfully requested.

For all the foregoing reasons, Applicant believes that claims 1-12, 14-16, 18, and 20-39, as amended, are patentable over the cited prior art and in condition for allowance. Reconsideration of the rejections and allowance of all pending claims 1-12, 14-16, 18 and 20-39 are respectfully requested.

Respectfully submitted,

WEBB ZIESENHEIM LOGSDON  
ORKIN & HANSON, P.C.

By 

Richard L. Byrne  
Registration No. 28,498  
Attorney for Applicant  
700 Koppers Building  
436 Seventh Avenue  
Pittsburgh, PA 15219-1818  
Telephone: (412) 471-8815  
Facsimile: (412) 471-4094  
E-mail: [webblaw@webblaw.com](mailto:webblaw@webblaw.com)